

Crime Mapping Using Alteryx

Methodology

alteryx

Table of contents

1	Introduction	3
2	Data Input.....	4
3	Lower Layer Super Output Areas (LSOA) Area Mapping	6
4	Crime Type Mapping.....	7
5	Filling in Zeroes.....	8
6	Time Series.....	9
7	Validation.....	11
8	Exporting the Data to Qlik Sense	13
9	Demographics Data.....	14
10	Cluster Analysis	16
11	K-Centroids Cluster Analysis	21
12	Crime App – Qlik Sense	22
13	Alteryx Tools.....	24
14	Data Sources.....	25

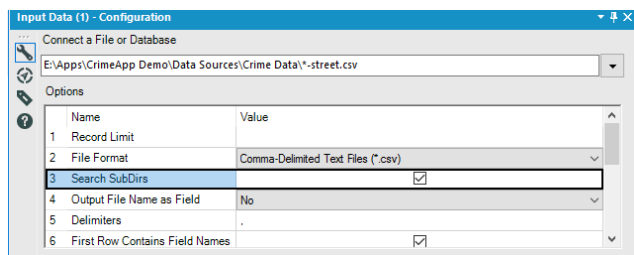
1 Introduction

This document provides the technical notes that go with our webinar. The objective of the project was to use a series of Alteryx workflows and predictive models to generate predicted crime figures for metropolitan areas in England over the next three years, and validate the data to test the quality of our models using data for the first three months of 2018 .

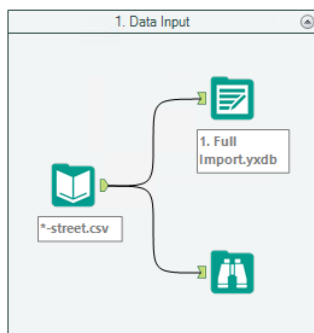
2 Data Input

The first step is to load all the crime data [1] into Alteryx. There are two different options that can be used : either load the data via the API provided or download all the data directly in a zipped file format. We have opted for the latter, but the former will be shown in a future edition.

We downloaded and unzipped the corresponding csv files only to find that each file for each location was in a separate subfolder. This would usually involve the use of the search function in windows explorer and some manual work, but with Alteryx we can use the wildcard function to select the appropriate files and simply select the option that searches through each of the subdirectories.



Now, all the files that end with the suffix ‘-street.csv’ will be imported. There are many files in the dataset, so at the first stage it is best practice to load it into an Alteryx database file format (.yxdb) which is the most efficient file type for reading and writing in Alteryx. This will speed up processing and we can create other outputs, including other Alteryx database files, once we have made our transformations. We name the file ‘1. Full Import.yxdb’.


















We also added a ‘Browse’ tool to see the data in some more detail. We can see that the key fields to us are “Crime ID”, “Month”, “LSOA Name” and “Crime Type”.

Record #	Crime ID	Month	Reported by	Falls within	Longitude	Latitude	Location	LSOA code	LSOA name	Crime type	Last outcome category
23465456	a4230f00...	2014-11	Essex Police	Essex Police	0.918776	51.893455	On or near Supermarket	E01021704	Colchester 022F	Shoplifting	Unable to prosecute susp...
23465457	353db0b9a...	2014-11	Essex Police	Essex Police	0.918776	51.893455	On or near Supermarket	E01021704	Colchester 022F	Shoplifting	Investigation complete...
23465458	62a568b94...	2014-11	Essex Police	Essex Police	0.918776	51.893455	On or near Supermarket	E01021704	Colchester 022F	Shoplifting	Court result unavailable
23465459	6d73a6a4...	2014-11	Essex Police	Essex Police	0.918776	51.893455	On or near Supermarket	E01021704	Colchester 022F	Shoplifting	Investigation complete...
23465460	249c724fb...	2014-11	Essex Police	Essex Police	0.926662	51.894320	On or near Supermarket	E01021704	Colchester 022F	Shoplifting	Investigation complete...
23465461	d46952bce...	2014-11	Essex Police	Essex Police	0.918776	51.893455	On or near Supermarket	E01021704	Colchester 022F	Vehicle crime	Investigation complete...
23465462	[Null]	2014-11	Essex Police	Essex Police	0.944599	51.892227	On or near Heatley Way	E01021705	Colchester 022G	Anti-social behaviour	[Null]
23465463	[Null]	2014-11	Essex Police	Essex Police	0.938092	51.895186	On or near Hawthorn Avenue	E01021705	Colchester 022G	Anti-social behaviour	[Null]
23465464	[Null]	2014-11	Essex Police	Essex Police	0.942002	51.896303	On or near Salary Close	E01021705	Colchester 022G	Anti-social behaviour	[Null]
23465465	2204ed9a0...	2014-11	Essex Police	Essex Police	0.950042	51.892765	On or near Fulmar Close	E01021705	Colchester 022G	Burglary	Investigation complete...
23465466	7203b8af2...	2014-11	Essex Police	Essex Police	0.942002	51.896303	On or near Salary Close	E01021705	Colchester 022G	Criminal damage and a...	Investigation complete...
23465467	5a49520ef...	2014-11	Essex Police	Essex Police	0.949562	51.889872	On or near Dunmoo Way	E01021705	Colchester 022G	Vehicle crime	Investigation complete...
23465468	853acce12...	2014-11	Essex Police	Essex Police	0.938092	51.895186	On or near Hawthorn Avenue	E01021705	Colchester 022G	Violence and sexual off...	Unable to prosecute susp...
23465469	1a7805871...	2014-11	Essex Police	Essex Police	0.949562	51.889872	On or near Dunmoo Way	E01021705	Colchester 022G	Violence and sexual off...	Unable to prosecute susp...
23465470	[Null]	2014-11	Essex Police	Essex Police	0.926662	51.888007	On or near Tabor Road	E01021706	Colchester 022H	Anti-social behaviour	[Null]
23465471	[Null]	2014-11	Essex Police	Essex Police	0.921379	51.889874	On or near Kerry Court	E01021706	Colchester 022H	Anti-social behaviour	[Null]

Below is a snippet taken from the results output when running the short workflow that can be seen above.

Results - Workflow - Messages

	 0 Errors	 164 Conv Errors	 0 Warnings	 1 Messages	 3880 Files	All
	Input Data (2)	4960 records were read from "E:\Apps\CrimeApp Demo\Data Sources\Crime Data\2018-03-suffolk-street.csv"				
	Input Data (2)	7616 records were read from "E:\Apps\CrimeApp Demo\Data Sources\Crime Data\2018-03-surrey-street.csv"				
	Input Data (2)	11412 records were read from "E:\Apps\CrimeApp Demo\Data Sources\Crime Data\2018-03-sussex-street.csv"				
	Input Data (2)	15216 records were read from "E:\Apps\CrimeApp Demo\Data Sources\Crime Data\2018-03-thames-valley-street.csv"				
	Input Data (2)	4225 records were read from "E:\Apps\CrimeApp Demo\Data Sources\Crime Data\2018-03-warwickshire-street.csv"				
	Input Data (2)	9748 records were read from "E:\Apps\CrimeApp Demo\Data Sources\Crime Data\2018-03-west-mercia-street.csv"				
	Input Data (2)	23088 records were read from "E:\Apps\CrimeApp Demo\Data Sources\Crime Data\2018-03-west-midlands-street.csv"				
	Input Data (2)	27674 records were read from "E:\Apps\CrimeApp Demo\Data Sources\Crime Data\2018-03-west-yorkshire-street.csv"				
	Input Data (2)	4600 records were read from "E:\Apps\CrimeApp Demo\Data Sources\Crime Data\2018-03-wiltshire-street.csv"				

3 Lower Layer Super Output Areas (LSOA) Area Mapping

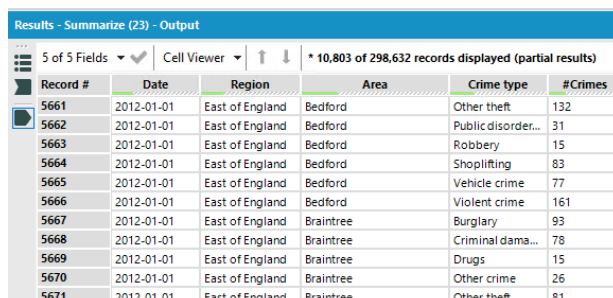
Now that the data has been loaded into an Alteryx database format, we can work more efficiently with our data. The first set of transformations is to remove the redundant information in the dataset and map the regions of England that we want to use in our final analysis to the LSOA areas listed in the crime dataset. LSOAs are a geographic hierarchy designed to improve the reporting of small area statistics within England and Wales, though we are just using English areas in our sample.

Below is the workflow that we used to do these transformations. Initially there are two streams; the top stream which transforms the original raw data (in the .yxdb format we created in step one) by removing all redundant information; and the second stream which maps each area to its corresponding region.

The first stream simply loads the raw data in the .yxdb, removes any rows that do not have valid crime IDs or LSOA codes and creates a new field which states the actual area version (an area represented by English Districts, Unitary authority's and London Boroughs. For instance, the LSOA area of "Mansfield 0131" and "Newark and Sherwood 001A" become "Mansfield" and "Newark and Sherwood" respectively.

The second stream loads in the 'Local Authority District to Region Lookup in England' file which has a list of all areas and its corresponding regions within England (where a region is one the following: Greater London, South East, South West, West Midlands, North West, North East, Yorkshire and the Humber, East Midlands or East of England). Before joining these two streams, we need to create a lookup to change some area names to directly match the area names we created in the new field in the first stream. For example, we need to change all instances of "Bristol, City of" and "Kingston upon Hull, City of" to "Bristol" and "Kingston upon Hull". To do this we use a 'Text Input' tool to find and replace the affected area names using the 'Find Replace' tool. We can now take a copy of the new area names in the form of an Alteryx database file in case we want to use it against other datasets.

Now join the two streams together using the 'Join' tool on the new "Area" field, convert original string format into date format for Alteryx using the 'DateTime' parsing tool and grouping the data into a useable format using the 'Summarise' tool.



The screenshot shows the 'Results - Summarize (23) - Output' window in Alteryx. It displays a table with 6 columns: Record #, Date, Region, Area, Crime type, and #Crimes. The table contains 11 rows of data, all dated 2012-01-01. The regions are 'East of England', and the areas are 'Bedford' and 'Braintree'. The crime types include 'Other theft', 'Public disorder...', 'Robbery', 'Shoplifting', 'Vehicle crime', 'Violent crime', 'Burglary', 'Criminal dama...', 'Drugs', 'Other crime', and 'Other theft'. The number of crimes per record ranges from 15 to 132.

Record #	Date	Region	Area	Crime type	#Crimes
5661	2012-01-01	East of England	Bedford	Other theft	132
5662	2012-01-01	East of England	Bedford	Public disorder...	31
5663	2012-01-01	East of England	Bedford	Robbery	15
5664	2012-01-01	East of England	Bedford	Shoplifting	83
5665	2012-01-01	East of England	Bedford	Vehicle crime	77
5666	2012-01-01	East of England	Bedford	Violent crime	161
5667	2012-01-01	East of England	Braintree	Burglary	93
5668	2012-01-01	East of England	Braintree	Criminal dama...	78
5669	2012-01-01	East of England	Braintree	Drugs	15
5670	2012-01-01	East of England	Braintree	Other crime	26
5671	2012-01-01	East of England	Braintree	Other theft	81

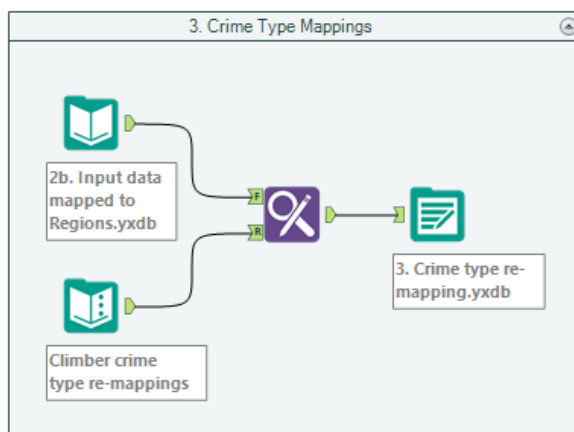
The last step in this stage is to load the transformed data shown above into a new Alteryx database file '2b. Input data mapped to Regions.yxdb'.

4 Crime Type Mapping

The next stage is a short one that groups existing crime types to our new crime groups for our final analysis. This is to counter the effects of changes in crime recording; such as reclassifications associated with legal enquiries and modern updates that happen over the years, for example new offences such as 'causing serious injury by dangerous driving' were added to the Homicide crime type in April 2013 or Action Fraud taking over the recording of fraud offences on behalf of individual police forces.

Create a 'Text Input' tool with our new mapping and use the 'Find Replace' tool to replace the relevant field from the '2b. Input data mapped to Regions' file we created at the end of the last stage. To finish off this stage we create a new Alteryx database file titled '3. Crime type re-mapping'.

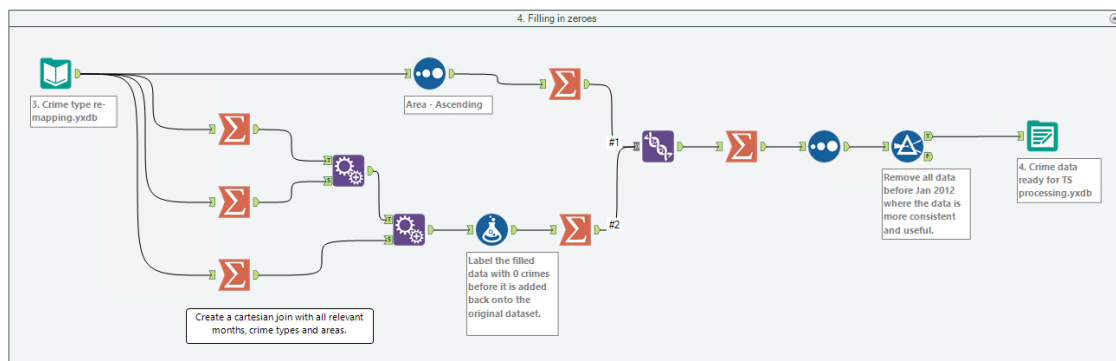
Crime type	Climber Crime Grouping
Anti-social behaviour	Non-violent crime
Bicycle theft	Theft
Burglary	Theft
Criminal damage and arson	Violent crime
Drugs	Non-violent crime
Other crime	Non-violent crime
Other theft	Theft
Possession of weapons	Violent crime
Public disorder and weapons	Violent crime
Public order	Non-violent crime
Robbery	Theft
Shoplifting	Theft
Theft from the person	Theft
Vehicle crime	Non-violent crime
Violence and sexual offences	Violent crime
Violent crime	Violent crime



5 Filling in Zeroes

Before we can forecast our data, we need to ensure that there are no missing fields. Reducing the number of crime groups in the last stage made it more unlikely that some areas didn't have a single crime in a certain group, playing around with the data we found that there are some instances where no crime was recorded in a particular month. We need to add a 0 in the rows where a crime ID was not present.

There are many ways to do this final transformation, one of which is shown in the workflow below where we have opted to use a cartesian join. We have used three different 'Summarise' tools each listing the distinct values for "Date", "Area and Region", and "Crime Type" in our data. We join these together in two steps using two "Append Fields" tools and assigning each with a value of 0 crimes. We then use a summarise tool with the '3. Crime type re-mapping.yxdb' data imported to create a similar data structure format to the cartesian join we just processed. The "Union" tool is used to combine the two streams together, essentially adding one set on top of the other. We then use another summarise tool to group the data so we are left with a dataset containing all the original data, and new rows for areas where there were no associated crime IDs (with a new crime count of 0 for each of those rows).



6 Time Series

Now we have reached the forecasting stage. All the transformations to our data have been done so we are now able to process it using the “Time Series” tools that Alteryx has. Before we use the tools, we take out the 2018 data (currently only January, February and March are available) which we will use to check against our forecast.

There are two options, to either run our time series using the “ARIMA” (Auto-Regressive Integrated Moving Average) tool or the “ETS” tool (Exponential Smoothing). At a high level, “ARIMA” utilises the moving average for its forecasting and can estimate a time series model either as a univariate model or one that includes covariates (additional fields which serve as predictors). ETS estimates a time series forecast by using a weighted average on past observations. Increasing weight is given to more recent values. ETS is also able to account for three different components of time series forecasting: seasonality, level and trend. Both ARIMA and ETS have their individual benefits for model suitability (detailed differences will be found in a future post), and we ran both models to test their suitability for our crime groups.

As mentioned, the native tools we would use are the “ARIMA” and “ETS” to run the time series and then the “TS Forecast” for the forecasting, but since we want to run these for more than one group we need to use the “TS Model Factory” and “TS Forecast Factory” which can be found in the Alteryx gallery (<https://gallery.alteryx.com>).

The “TS Model Factory tool” first estimates the time series model for multiple groups using either “ARIMA” or “ETS” as configured. If using “ARIMA” you can use covariates (predictors) for the model estimation. You can configure a target field to forecast and the grouping field. You can input the time period type depending on whether your data is monthly or hourly for instance and can define a set starting period. The “TS Forecast Factory” tool then provides the forecasts for your models for a user-specified number of future periods. It allows you to configure different confident level percentages to give you upper and lower confidence bounds. For each confidence level, the expected probability that the true value will fall within the provided bounds corresponds to the confidence level percentage. The confidence level describes a level of uncertainty associated with the interval estimate, so using a typical confidence interval of 95% we can say that using the same method of modelling for a different sample set would yield the results to fall within the interval estimates 95% of the time.

The process for running either version is essentially the same when using the “TS Model Factory”, but involves a different configuration for each.

TS Model Factory (2) - Configuration

Model Configuration

Type of Time Series Model

☐ ARIMA...

☒ ETS

Select the target field

#Crimes

Select the grouping field

Key

Time period type

☐ Hourly

☐ Daily (all days)

☐ Daily (weekdays)

☐ Weekly

☒ Monthly

☐ Quarterly

☐ Annually

☐ Other

☒ Series starting period (optional)

The year the series starts

2012

The week, month (numeric), or quarter of the series start

1

TS Forecast Factory (3) - Configuration

Configuration

The field name for the point forecast

ETS forecast

The percentage value of the larger confidence interval

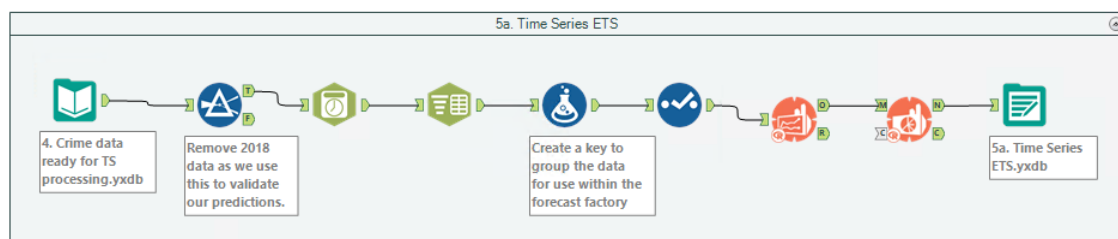
95

The percentage value of the smaller confidence interval

80

The number of future periods to forecast (assuming no covariates in model)

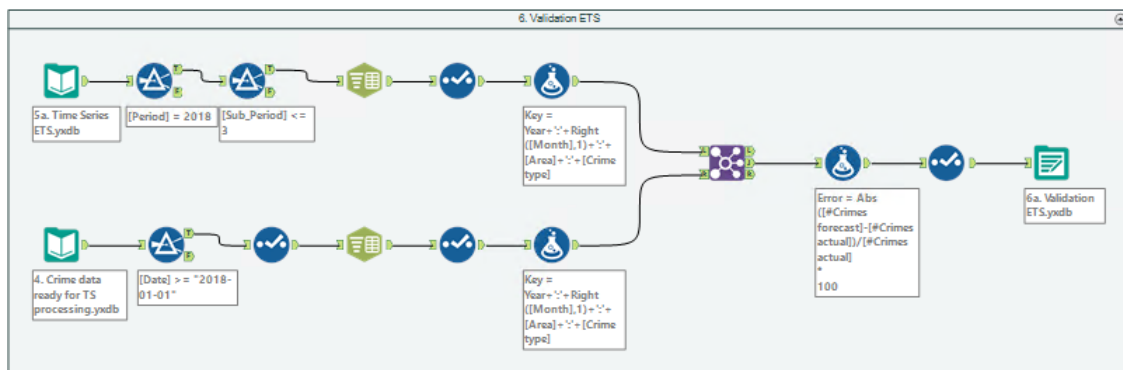
6



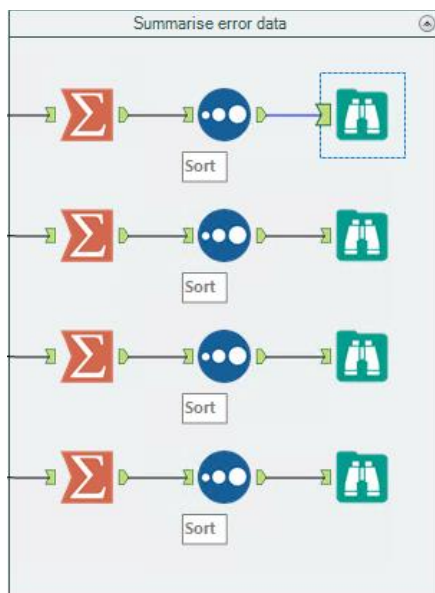
7 Validation

Having run both “ARIMA” and “ETS” models, allows us to compare our forecasted data to real life data for 2018, which was excluded from our forecasting. There are two streams in our workflow; the top workflow pulling in ‘5a. Time Series ETS.yxdb’ which was created in the last stage, and ‘4. Crime data ready for TS processing.yxdb’ which was the historical crime data that we transformed and saved as an Alteryx data file. Both are then reduced to include the first three months of 2018 and joined using a new key so we can calculate the error between the actual and forecasted values.

We run this workflow twice, once for “ETS” forecasted data and once for “ARIMA” forecasted data.



Once our workflows have run, we summarise the data to see how the forecast did against different groups.



Errors across the different regions:

Record #	Region	Avg_Error
1	East Midlands	13.243374
2	East of England	11.625746
3	London	8.067116
4	North East	9.832644
5	North West	11.957021
6	South East	12.376038
7	South West	15.177387
8	West Midlands	13.486246
9	Yorkshire and The Humber	10.77419

Error against our “ETS” model was shown to be 12.2%, suggesting our forecasts are 87.8% accurate:

Record #	Avg_Error
1	12.204109

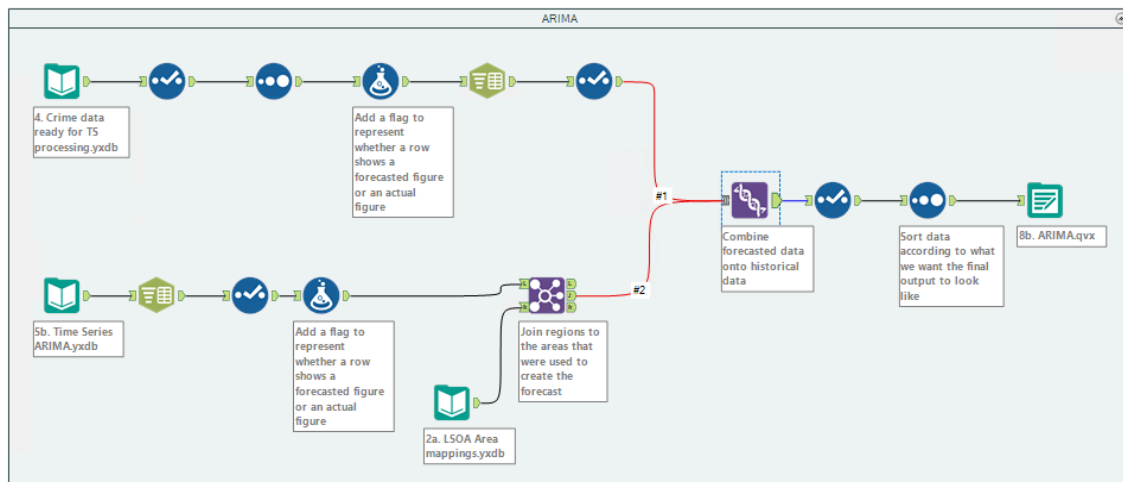
Error against our “ARIMA” model was shown to be 12.1%, suggesting our forecasts are 87.9% accurate:

Record #	Avg_Error
1	12.103671

Since we found that the “ARIMA” model performed slightly better than our “ETS” model, we decide that we will use that data for our analysis.

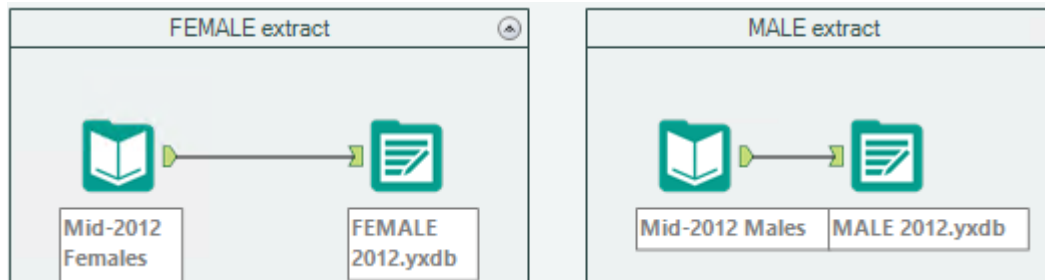
8 Exporting the Data to Qlik Sense

We need to produce a single table with the data that we need to conduct an in-depth analysis of crime using Qlik Sense. To do this we pull in the same Alteryx data files as we did in the last stage, '5a. Time Series ETS.yxdb' and '4. Crime data ready for TS processing.yxdb' (note we are not using a validation set as that was more to understand our models better). The first stream pulls in the transformed historical data and adds a new flag to indicate that it is historic data. The second stream pulls in the forecasted data and adds a flag to indicate that it is forecast data. Before we combine the two streams, we import the "2a. LSOA Area mappings.yxdb" file back to map regions back to areas, which we previously lost whilst performing the time series model as it can only be used to group against a single field. Once the data is combined we use a "Select" tool to rename and reorder our fields and then a "Sort" tool to sort our data as necessary. We then use an "Output" tool to indicate that we would like to output the data as a QlikView data exchange format (.QVX) as described in Alteryx to enable an efficient import into Qlik Sense.



9 Demographics Data

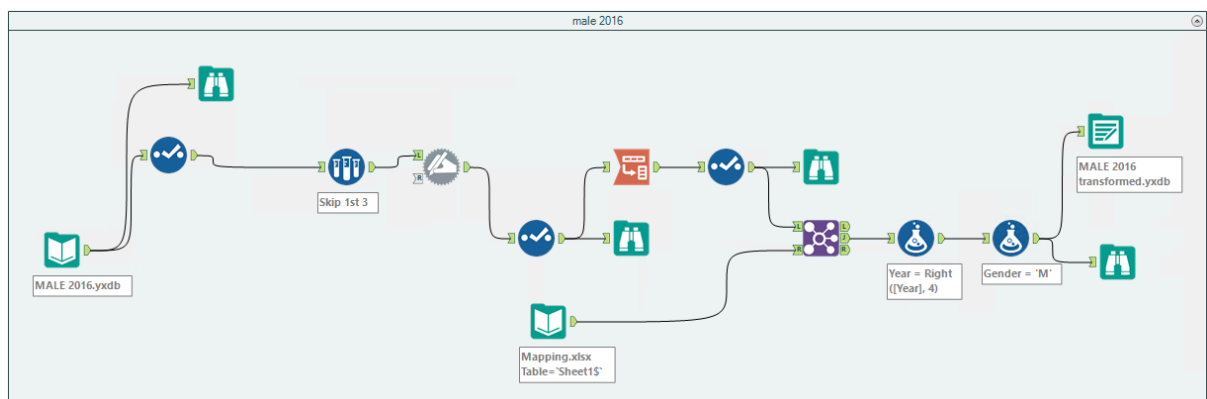
To enrich our crime analysis, we have included an additional dataset. We used demographic data for England, broken down by age and sex for each LSOA. Using Alteryx, we created workflows to extract data for 2012 to 2016. We use the 2016 data to estimate each year onwards as that data has not yet been released. Similar to the approach we took for the crime data, our first stage involved the use of imported the raw data for each year and creating an Alteryx data file for more efficient transformations at the next stage. Unfortunately, since the data was different in each sheet, we would have to use multiple files.



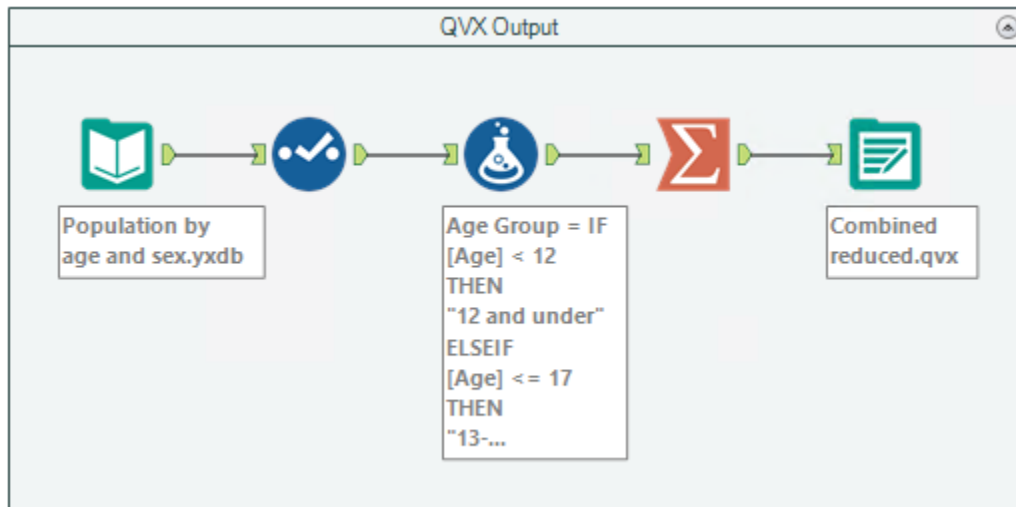
Below is an example of the raw data. We know this isn't a simple flat file and it's not ideal for analysis as it is, especially as we want to enhance our crime forecasts with it. Fortunately for us, Alteryx is proficient at data preparation and can quickly build a workflow to transform the data that's useable.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Mid-2015 Lower Layer Super Output Area population estimates for England and Wales (Supporting Information)														
2	Single Year of Age, Males														
3															
4	Area Codes	Area Names	All Ages	0	1	2	3	4	5	6	7	8	9	10	
5	E06000047	County Durham	255,299	2,715	2,890	2,994	2,985	3,104	3,014	2,917	2,986	2,955	2,834	2,780	
6	E01020634	County Durham 001A	745	15	11	10	11	12	7	6	8	9	6	8	
7	E01020635	County Durham 001B	649	3	8	13	2	8	4	8	8	5	4	4	
8	E01020636	County Durham 001C	864	13	6	7	18	16	15	16	12	11	14	6	
9	E01020654	County Durham 001D	921	9	14	15	8	21	13	9	11	15	10	9	
10	E01020676	County Durham 001F	738	9	6	8	7	11	6	8	8	3	7	6	

We load in the Alteryx database files we just created, and using the "Select" tool remove any fields we don't need. We then use the "Sample" tool to remove the first three rows followed by the "Dynamic Rename" tool to embed the labels as our field headers. We now need to transpose the data from row D onwards to ensure it catches all the ages in the format needed. We join a new mappings file to ensure to get the correct areas based on each LSOA which we join the demographic data to, and finally create two new fields defining the year and gender, and creating another Alteryx database file for each input file.



Now to tidy up the table. Group the ages according to age bands required and reduce the data into a useable number of rows by using the “Summarise” tool, before exporting to the .QXV format for Qlik Sense. We also take the Alteryx database file version to use later.



10 Cluster Analysis

To further enrich the data, we add another element to our crime forecasts using predictive grouping. We use the Indices of Deprivation 2015 [6] to provide a set of relative measures of deprivation for small areas (Lower-layer Super Output Areas) across England, based on seven different domains of deprivation:

- Income Deprivation
- Employment Deprivation
- Education, Skills and Training Deprivation
- Health Deprivation and Disability
- Crime
- Barriers to Housing and Services
- Living Environment Deprivation

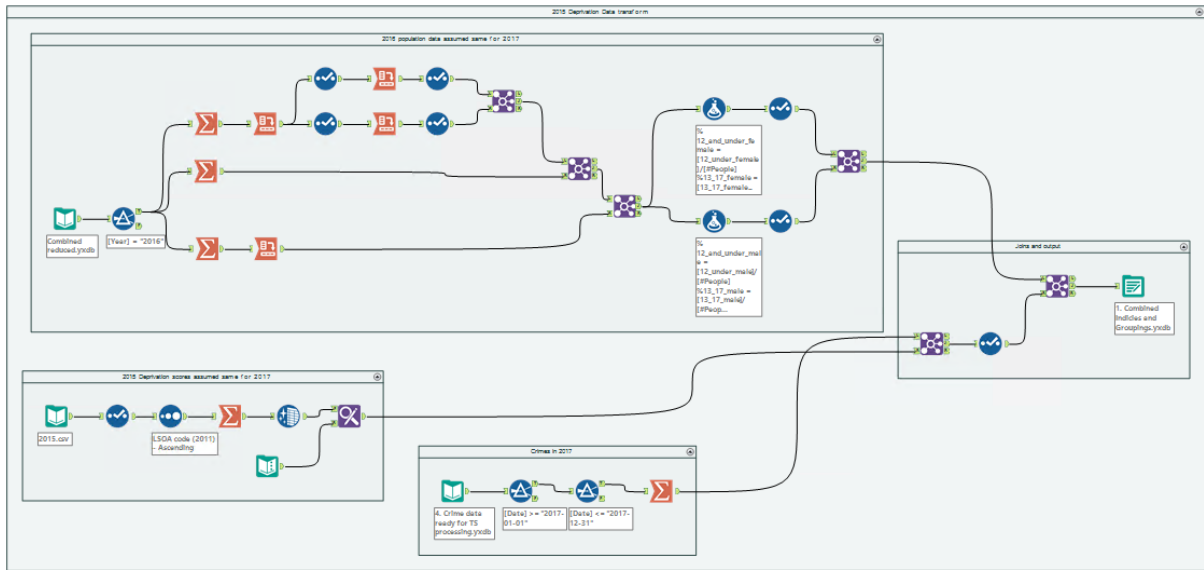
Each of these domains are based on a basket of indicators. As far as possible, each indicator is based on data from the most recent time point available. For our purposes, we use the latest data from the English indices of deprivation 2015 and use them for the following years.

We take the indicators for each domain, excluding crime as that's what we are forecasting, and use it for our clustering analysis. To do this effectively, we need to find out which of these factors are more likely to impact crime and exclude those that do not.

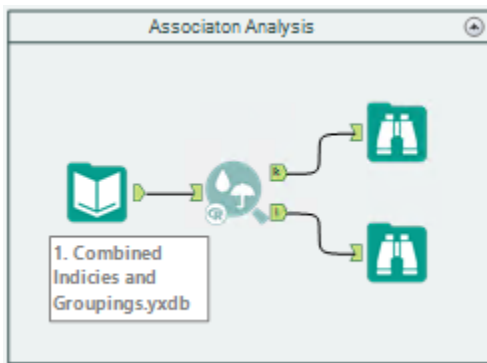
Before we do this, we need to join up the appropriate data to do this using some of the Alteryx database files we have used previously. The first stream imports 'Combined reduced.yxdb' which comes from the previous stage and has the population data by area for each age group and gender.

We then use a series of "Cross Tab" and "Summarise" tools to create fields detailing the number of people of each age group and gender for each area. We use this to create new fields detailing these figures as a percentage and remove the absolute numbers.

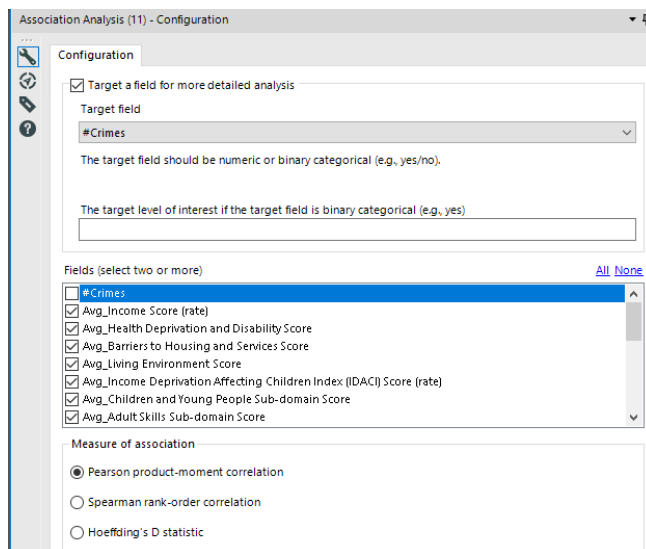
The second stream imports a new file detailing the 2015 Deprivation scores which we have assumed to be the same for 2017. We clean the data using a "Data Cleansing" tool and use a "Find Replace" tool in a similar fashion to stage 2 to ensure the area labels are consistent. We then join this data to our historical crime data file (4. Crime data ready for TS processing.yxdb) and join this new combined dataset to the first stream mentioned above. The result is a dataset that includes the number of crimes for each area, the indicators for each deprivation domain as listed previously and the percentage of each age group by gender, which we export as an Alteryx database file called 'Combined Indices and Groupings.yxdb'.



Association Analysis

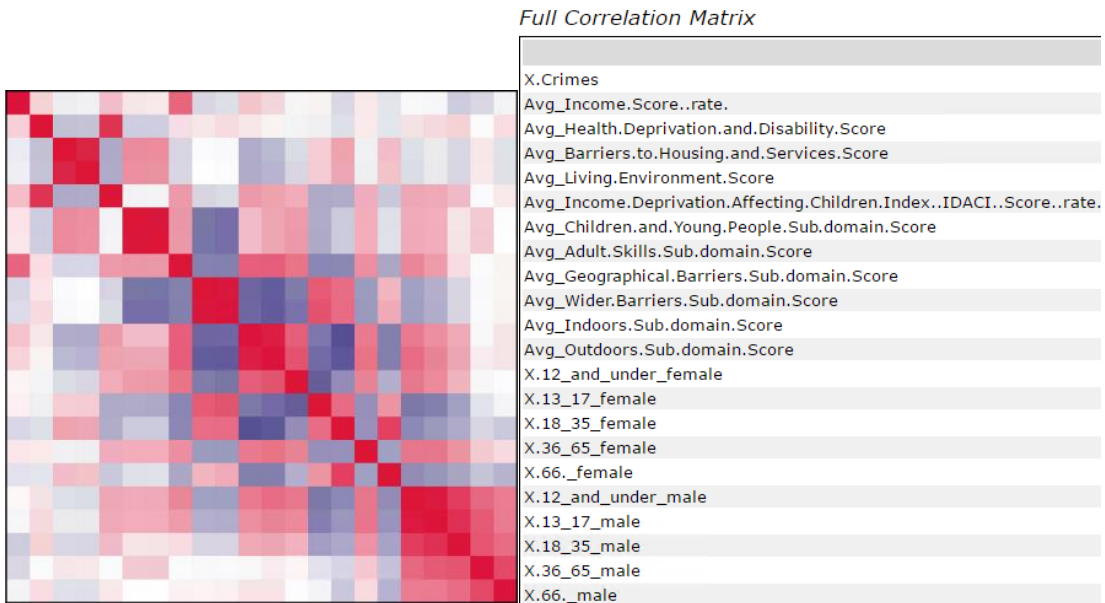


The next stage is to understand which of those fields has the biggest effect on the number of crimes in each area. We can find this easily using the “Association Analysis” tool:



By selecting the target field as the number of crimes, we can use the remaining fields to determine whether there is a (bivariate) association between those fields and the number of crimes.

The output shows the following, where the left image is a correlation matrix (blue represents -1 and red represents +1) and the right image details the actual correlation values.



We have decided to use values over 0.5 which suggests a good correlation with crime numbers moving forward for our clustering analysis. This leaves us with the following fields:

- Income score rate
- Income deprivation affecting children index
- Outdoors score
- % of people age 18 to 35 female
- % of people age 18 to 35 male

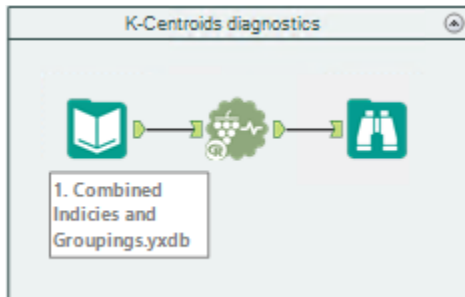
There is an option here to reduce the number of variables from the five above further if we like. The benefit of having lower, more correlated variables to use is that they have more control over how the model performs. An excessive number of variables can easily result in decreased model accuracy. Because of this, we have decided to use

- Contents
- No table of contents entries found.
- the two variables which show the greatest correlation.
- K-Centroids diagnostics

It should be mentioned that predictive grouping is a topic that is worth following more in its own right and we briefly describe how it is used in this section, with the intention of a possible future post where we can discuss it in greater detail.

K-Centroids represent a class of algorithms for doing what is known as partitioning cluster analysis. These methods work by taking the records in a database and dividing (partitioning) them into the “best” K groups based on some criteria. Nearly all the partitioning cluster analysis methods accomplish their objective by basing cluster membership on the proximity of each record to one of K points (or “centroids”) in the data. The objective of these clustering algorithms is to find the location

of the centroids that optimises some criteria with respect to the distance between the centroid of a cluster and the points assigned to that cluster for a pre-specified number of clusters in the data. [7]



At a high level, the “K-Centroids Diagnostic” tool is designed to allow the user to assess the appropriate number of clusters to specify given the data and the selected clustering algorithm (K-Means, K-Medians, or Neural Gas). We use the fields listed above, select standardisation to standardise the scores, select a clustering method (we will use K-Means in this example) and select our bound of clusters to consider using the minimum and maximum options.

The screenshot shows the "K-Centroids Diagnostics (23) - Configuration" window. It has two tabs: "Configuration" (selected) and "Graphics Options".

Fields (select two or more) (All None):

- ☐ #Crimes
- ☒ Avg_Income Score (rate)
- ☐ Avg_Health Deprivation and Disability Score
- ☐ Avg_Barriers to Housing and Services Score
- ☐ Avg_Living Environment Score
- ☒ Avg_Income Deprivation Affecting Children Index (IDAC) Score (rate)
- ☐ Avg_Children and Young People Sub-domain Score
- ☐ Avg_Adult Skills Sub-domain Score

Standardize the fields:

- ☒ z-score
- ☐ Unit interval

Clustering method:

- ☒ K-Means
- ☐ K-Medians
- ☐ Neural Gas

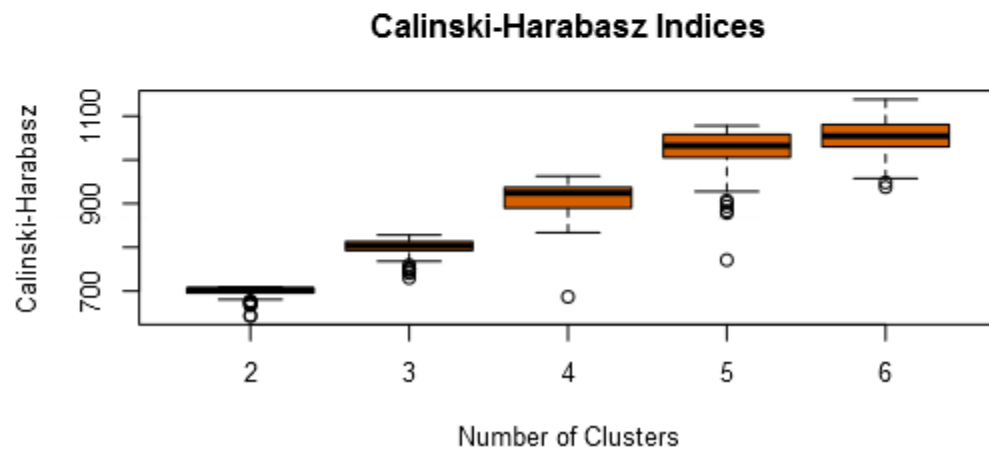
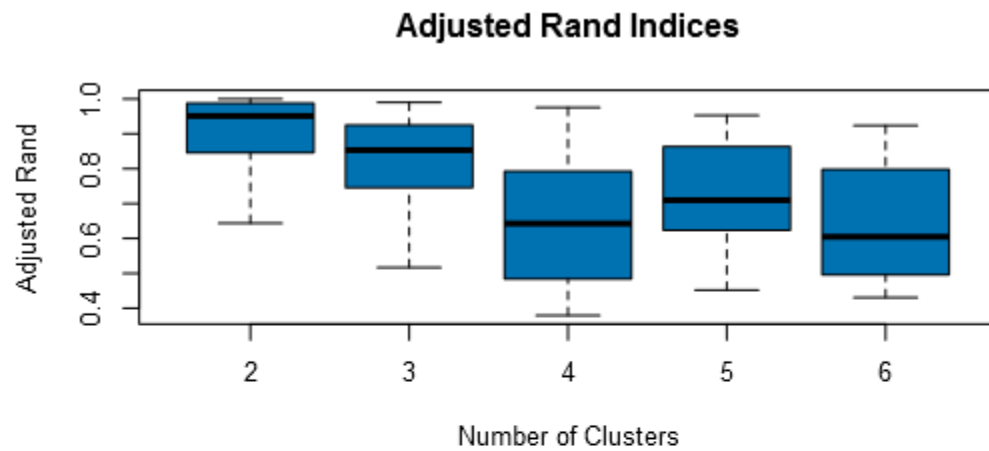
Minimum number of clusters: 2

Maximum number of clusters: 6

Bootstrap replicates: 50

Number of starting seeds: 3

After running the tool, we can see the following charts in the browse tool. We can use both indices to determine the number of clusters that we should use.



Simply, the preferred number of clusters based on each measure corresponds to one with the highest mean and median of the solutions compared. We can choose the number of clusters that seems like the best fit to this condition. As we would like to group our areas by more than two clusters, an observation that we follow with is that five clusters have a reasonably high mean in the Rand Index and the highest mean in the Calinski–Harabasz index when compared to the other options. It is therefore reasonable to suggest that using five clusters is the way forward.

11 K-Centroids Cluster Analysis

Now that we know the number of clusters that we want to use, we can use the “K-Centroids Cluster Analysis” tool. Note that we are still using the ‘Combined Indices and Groupings.yxdb’ file as we have not made any transformations.

First, we “K-Centroids Cluster Analysis” tool to our data and we notice that the configuration is very similar to that of the “K-Centroids Diagnostic” tool, except instead of selecting an upper and lower cluster bound we now input the number of clusters that we want instead.

K-Centroids Cluster Analysis (1) - Configuration

Configuration Plot Options Graphics Options

Solution name
Crime clusters

Fields (select two or more) [All](#) [None](#)

- ☐ #Crimes
- ☒ Avg_Income Score (rate)
- ☐ Avg_Health Deprivation and Disability Score
- ☐ Avg_Barriers to Housing and Services Score
- ☐ Avg_Living Environment Score
- ☒ Avg_Income Deprivation Affecting Children Index (IDACI) Score (rate)
- ☐ Avg_Children and Young People Sub-domain Score
- ☐ Avg_Adult Skills Sub-domain Score

☒ Standardize the fields

☒ z-score
☐ Unit interval

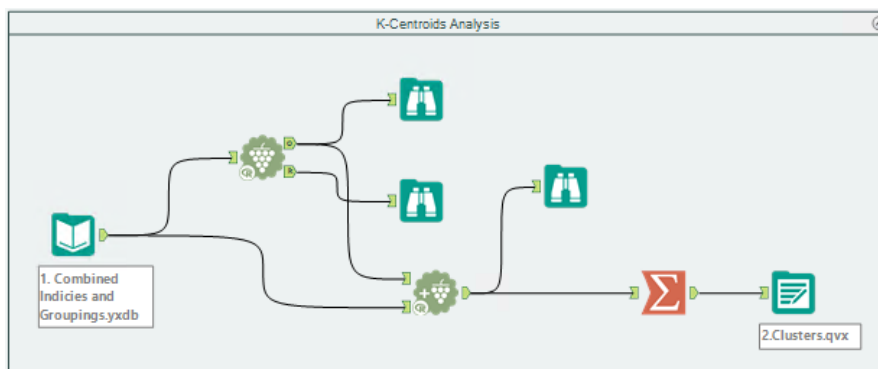
Clustering method

☒ K-Means
☐ K-Medians
☐ Neural Gas

Number of clusters
5

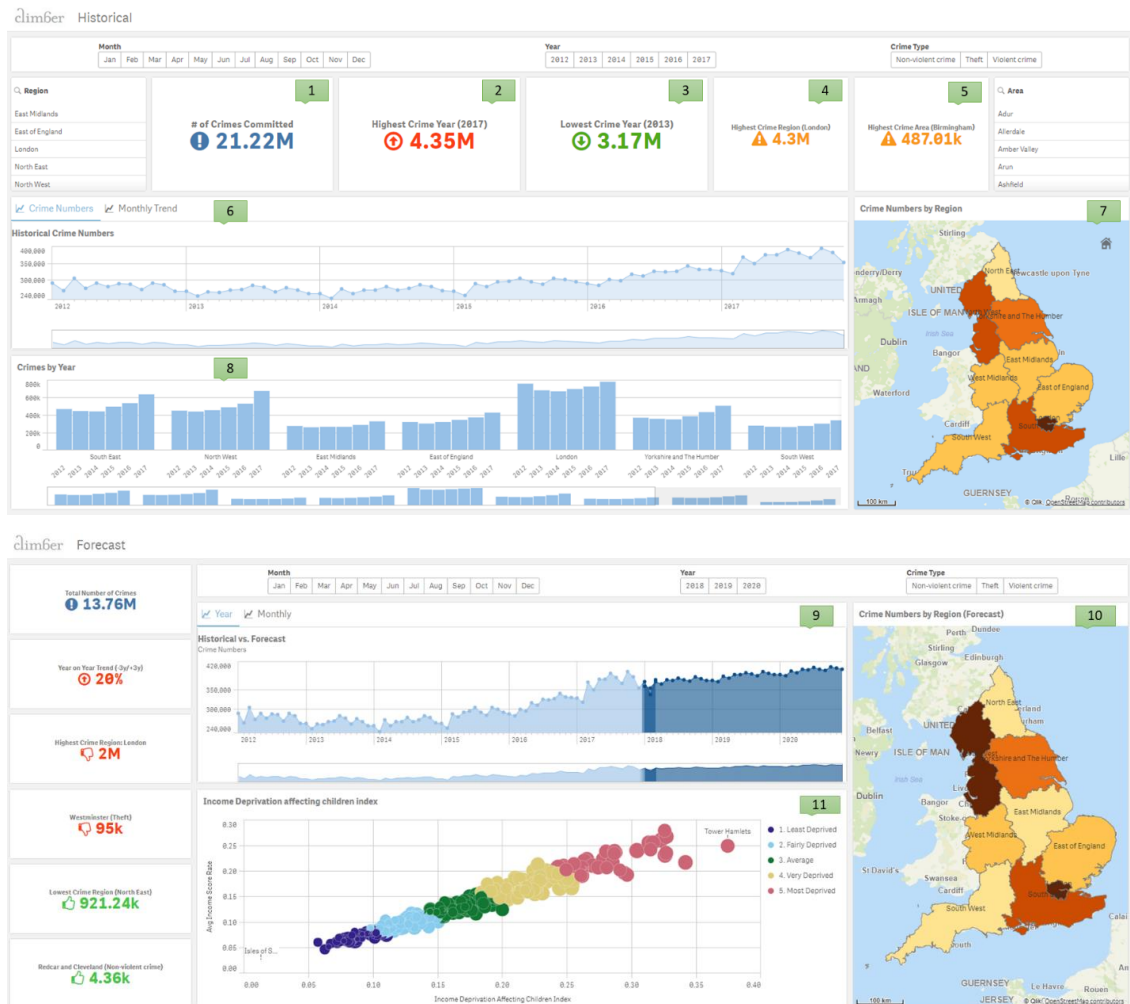
Number of starting seeds
10

This will create a field that assigns each area to a cluster numbered between 1 and 5 based on the algorithm. We then attach this data to the “Append Cluster” tool which appends the field to the original data that we used, so there is a new column which indicates the cluster group for each area alongside the rest of the data. We then use the “Summarise” tool to regroup and remove the information we no longer need (the predictor fields) before exporting the file in the .QVX format for Qlik Sense.



12 Crime App – Qlik Sense

We loaded in all the relevant files into Qlik Sense including the KML files for the maps ([3],[4]) and below you will find the two screens that we have created. The first showcasing our initial results and the second a presentation of our forecasting.



Visualisation Number	Description
1	The total number of crimes committed between January 2012 and December 2017
2	The year which contained the highest number of crimes
3	The year which contained the lowest number of crimes
4	The region with the highest number of crimes
5	The area with the highest number of crimes
6	The crime trend between January 2012 and December 2017

7	A map coloured by the number of historical crimes within each region. The darker colours represent a greater number of crimes
8	The total number of crimes each year between January 2012 and December 2017 for each region
9	The crime trend between January 2012 and December 2017 and the forecasted trend between January 2018 and December 2020. The darker strip represents the difference between our forecast for January – March 2018 compared to the real data
10	A map coloured by the number of forecasted crimes within each region. The darker colours represent a greater number of crimes
11	A scatter plot showing the average income score rate and the income deprivation affecting children index score for each area. The size of each bubble represents the number of crimes in each area and each colour represents its cluster group, from least deprived to most deprived

13 Alteryx Tools

Below is the complete list of the tools that was used within Alteryx

Tool Group	Tool
In/Out	Browse
	Input Data
	Output Data
	Text Input
Preparation	Data Cleansing
	Filter
	Formula
	Sample
	Select
	Sort
	Unique
Join	Append Fields
	Find Replace
	Join
	Union
Parse	DateTime
	Text to Columns
Transform	Cross Tab
	Summarise
	Transpose
Data Investigation	Association Analysis
Time Series	TS Model Factory
	TS Forecast Factory
Predictive Grouping	Append Cluster
	K-Centroids Cluster Analysis
	K-Centroids Diagnostics
Developer	Dynamic Input

14 Data Sources

[1] All crime data

<https://data.police.uk/data/>

[2] Local Authority District to Region (December 2016) Lookup in England

<http://geoportal.statistics.gov.uk/datasets/local-authority-district-to-region-december-2016-lookup-in-england>

[3] KML for areas - English Districts, UAs and London Boroughs, 2011

https://borders.ukdataservice.ac.uk/easy_download_data.html?data=England_lad_2011

[4] KML for regions - Regions (December 2017) Ultra Generalised Clipped Boundaries in England

<http://geoportal.statistics.gov.uk/datasets/regions-december-2017-ultra-generalised-clipped-boundaries-in-england>

[5] Demographics data - Lower Super Output Area Mid-Year Population Estimates

<https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/lowersuperoutputareamidyearpopulationestimates>

[6] English indices of deprivation 2015

<https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015>

[7] Alteryx reference manual

<https://help.alteryx.com/11.0/index.htm>